*Research Report*

# National Board for Professional Teaching Standards Bias-Reduction Training: Impact on Assessors' Awareness

E. Caroline Wylie

Michelle Y. Szpara

Research & Development

March 2004
RR-04-12

# National Board for Professional Teaching Standards Bias-Reduction Training: Impact on Assessors' Awareness

E. Caroline Wylie
ETS, Princeton, NJ

Michelle Y. Szpara
Long Island University, Brookville, NY

March 2004

**Abstract**

This study is an in-depth investigation of the National Board for Professional Teaching Standards (NBPTS) bias-reduction training, from the perspective of assessors. The research examined how successful the bias training was in guiding assessors to recognize their biases and to identify actions to be used to reduce the impact of bias on their scoring decisions. The study focused on portfolio video entries to allow for a greater exploration of bias, since assessors are provided with a visual and aural depiction of the candidates and their students.

Key words: National Board for Professional Teaching Standards (NBPTS), assessor training, bias reduction

**Acknowledgements**

The authors wish to extend thanks to the site supervisors, trainers, and assessors at each of the study sites who agreed to participate in the study, and to the reviewers of early drafts of this report. Any opinions expressed in the publication are those of the authors and not necessarily of ETS.

## Table of Contents

## List of Tables

Page

## List of Figures

Page

## Introduction

The National Board for Professional Teaching Standards (NBPTS) utilizes complex performance assessments to certify accomplished teachers. The validity of the NBPTS assessment process rests largely on the professional judgments of trained assessors, applying scoring rubrics in a fair and consistent manner to candidate responses. However, assessors have the potential to introduce construct-irrelevant variance into the scoring process. For example, assessors may view certain teaching styles more favorably, particularly styles similar to their own. In addition, assessors may have internalized societal views of certain subpopulations, such as those defined by race, class, or gender. Because these stereotypes generally exist below the level of conscious attention, deliberate attention is required to identify and screen out these messages (Devine, 1989; Brief of Amicus Curiae, 2003).

NBPTS recognizes the importance of minimizing sources of construct-irrelevant variance by including bias-reduction exercises in the assessor training. Bias-reduction training, as one aspect of the overall assessor training process, seeks to bring subconscious influences to the surface, and by doing so, minimize their effects on scoring. Assessors encounter four specific bias-reduction exercises, developed on principles of adult learners' needs and current research on the role of experiential learning and reflection. Two exercises focus on personal preferences and biases, one on societal biases, and one on writing biases.

While anecdotal evidence suggests these exercises may be effective in reducing bias, no formal research studies have been conducted to assess the effectiveness of the bias-reduction exercises in the overall context of the NBPTS assessor training. The present study was therefore designed to explore the impact of the bias-reduction exercises on assessors' understanding of bias. This study is part a larger investigation of the NBPTS bias-reduction training, from the perspectives of both trainers and assessors. The research presented in this paper focuses on data collected from assessors. Specifically, the research examines how successful the bias training was in guiding assessors to recognize their biases and to identify actions to reduce the impact of bias on their scoring decisions. The primary and secondary research questions addressed in this paper are as follows:

What changes, if any, are seen in assessors' awareness of their biases as a result of the training and scoring process? Specifically,

1. What types of bias admissions do assessors make—that is, how do assessors take ownership of their biases?
2. What examples of bias do assessors provide?
3. What changes are seen in admissions of bias and examples of bias over time? In other words, can progression over the course of the training be observed?
4. What sociolinguistic techniques are employed by assessors in the process of admitting biases and/or supplying examples of biases?
5. What potential actions do assessors describe that could help them make fair and accurate scoring judgments?
6. What is the relationship between assessors' admissions of bias in their "trigger lists" and essays?

The evidence used to respond to these questions comes primarily from essay prompts that assessors were asked to respond to three times during the training and scoring session. Where appropriate, essay-based evidence is supplemented by evidence from "trigger lists," which are personalized lists of potential triggers that the assessors develop over the course of the scoring session. Triggers include anything that evokes a biased response that could cause an assessor to award a higher or lower score than the candidate response deserved.

All data collection for this study took place during the Summer 2002 NBPTS assessor training for portfolio video-based entries. Video-based entries were used in order to allow for a greater exploration of bias, since assessors are provided with both a visual and aural depiction of the candidates and their students. In other portfolio entries, an assessor has only a written description and paper artifacts from the classroom.

Linguistic and thematic analyses were used to examine how assessors discussed bias – whether they were explicit in taking ownership of their own biases and in naming the biases that they held—and how assessors addressed ways to reduce bias. On several occasions during the NBPTS training protocol, the assessors were provided with lists of categories of biases, as well as examples of actions to take to reduce the impact of those biases. It was recognized that some assessors might largely adopt the bias-related words and phrases used in training as their own, whereas other assessors might personalize the information and actively apply it to their own situation. In either case, assessors had to choose which biases to include in their trigger lists and in their essay responses, providing some degree of personalization of the training material.

## Outline of NBPTS Training and Scoring

NBPTS assessments consist of ten entries whose tasks involve videotaping classroom lessons; submitting student work; writing commentaries to accompany video or student work; documenting accomplishments outside the classroom such as working with parents, professionals, or other colleagues; and responding to six constructed-response exercises that focus on content and content-based pedagogical knowledge. A significant aspect of the NBPTS certification process is the scoring of candidates' entries. Multiple scoring sites are established during the summer period. All responses to a particular entry are scored at the same site by a group of locally recruited assessors, all of whom are teachers, led by a lead trainer. Specific efforts are made to recruit diverse groups of assessors.

Assessors must complete a rigorous 4 day training period before they are allowed to score candidate performances. During the training process, the trainer works closely with assessors to ensure that they understand the requirements of the entry, how to review candidates' submissions, how to collect evidence, and how to make professional judgments using the entry-specific scoring rubric. Trainers are guided by a detailed written protocol that standardizes training across entries and certificates by providing talking points for each aspect of the training. In addition, trainers attend a session that gives them opportunities to both observe and participate in modeled training and to practice implementing it before they train assessors.

A significant portion of time is devoted during the training period to assessors reviewing benchmark and training cases—candidate responses that have been selected to represent the each of the rubric score points. Assessors review and score 12 to 14 benchmark and training cases. During the final day of training, assessors complete a qualifying round, where they independently score five to six prescored cases and must demonstrate a sufficient degree of accuracy in scoring and in the note-taking process that supports the score.

As part of the overall assessor training process, NBPTS currently addresses the need for bias-reduction training through a focus on potential "bias triggers." A trigger is something that an assessor identifies that may cause him or her to score a performance higher or lower than it deserves. Assessors are taught what to look for and score in a performance (through discussion of the rubrics, benchmark examples, etc.) and what to exclude from the overall evaluation (through the bias-reduction exercises). The summary below focuses on the four main components of the assessor training that relate to bias-reduction: two exercises designed to raise

assessors' awareness of their personal preferences, a third exercise concentrating on societal bias, and a fourth exercise on writing biases.

The first bias-reduction exercise, "Awareness of personal preferences," focuses on assessors' personal preferences by asking them to consider what indicates competence and incompetence when they meet people for the first time in an interview or other setting, and then in a school context. The initial part of the exercise is deliberately situated outside the context of teaching in order to help assessors recognize that making "snap judgments" is a process that humans engage in frequently. Assessors' responses regarding signals of competence typically range from teaching-related comments such as "knowledgeable about content" to other comments such as "professional dress" or "articulate." The role of the trainer at this stage is critical as he or she leads assessors to differentiate between comments that are grounded in the NBPTS Standards[1], and therefore legitimately part of the domain being assessed, and comments that indicate personal preferences. The goal is for assessors to recognize personal triggers that may make them favor or disadvantage a candidate based on information that is not pertinent to the scoring decision. This exercise helps bring previously unarticulated preferences to a conscious level.

During the second exercise, also focused on awareness of personal preferences, assessors are presented with a series of brief vignettes that describe a range of classroom situations. Assessors are asked to "finish the story"—that is, to write a brief description of what they think might happen next—and then share these stories with the larger group. The trainer's role is to help assessors realize how differently they each interpreted the various situations, and how easily they imposed their own experiences and understandings on the described situations. This exercise concludes with a discussion of the role of "reasonable inferences" as part of making a professional judgment about a candidate's response versus "fiction writing" which can occur when an assessor exceeds the bounds of reasonable inference based on the evidence presented.

The third exercise, "Awareness of societal bias," is conducted later in training when assessors have become more comfortable with one another, and directly addresses societal biases such as racism, sexism, and classism. This exercise does not attempt to correct "misinformation" about particular groups of people per se; instead, it seeks to demonstrate how assessors may have developed an extensive knowledge of the stereotypes that are prevalent in U.S. society today and are perpetuated widely in the media (van Dijk, 1987). First, assessors individually write word

associations in response to stimulus phrases such as, “Hispanic woman,” “urban,” “youthful,” etc. Next, assessors write a brief reflection on any patterns that they notice in their own responses, and consider what might have caused those patterns. The trainer then leads the group in a discussion, reassuring assessors that they do not need to share personal information. A strong emphasis of this exercise is that everyone has a reaction at some level to circumstances different than their own, and that often those reactions are influenced by stereotypes. Becoming aware of how stereotypes can unintentionally become part of one’s collective knowledge, and beginning to examine and interrogate these stereotypes, are important outcomes for assessors during this exercise.

In the fourth exercise, “Awareness of writing biases,” assessors review short excerpts of candidates’ writing and extrapolate what they think the rest of the response might be like. The excerpts are selected to exemplify a range of writing styles. After discussion of these excerpts, the trainer informs the assessors that all of the excerpts came from high-scoring NBPTS performances. The aim of this exercise is for assessors to identify their own writing style preferences and consider how these preferences can impact judgments. Writing bias, while not necessarily an obvious area of bias, is important given the format of the NBPTS assessment in which candidates present their approaches, rationales, analyses, and conclusions exclusively in writing.

At the beginning of the training, assessors establish a personal record of “triggers” that might make them score a candidate higher or lower than deserved. At the end of each bias exercise, assessors add to their lists, and as they watch videotapes and discover other triggers, they are asked to note those as well. Assessors keep their trigger lists throughout scoring and are expected to refer to them often to minimize the influence of bias on their scoring decisions.

## Literature

The validity of the scoring process for performance assessments relies in part on identifying and minimizing sources of construct-irrelevant variance. The current research draws upon research and theories in diverse fields, seeking to develop a comprehensive understanding of how assessors’ biases may interact with their ability to score fairly, and how assessors’ active examination of biases may reduce these interaction effects. The industrial psychology literature contributes much in terms of studies of performance appraisals, including training of raters,

effects of hidden biases, and methods for reducing the influence of biases. Cognitive and social psychology literature provides definitions of levels of prejudice or bias, as well as an understanding of why biases are so difficult to address. Research-based evidence can be found in the fields of cognitive and social psychology to support specific strategies for prejudice reduction, including those strategies utilized by the National Board in its assessor training. Finally, research from the field of education gives further credence to the importance of teachers recognizing societal biases they may have amassed in their collective knowledge, and the need for teachers to screen out the effects of such biases in their interactions with and evaluations of their students. These diverse fields of literature are further described below.

The industrial psychology literature contains extensive research on performance appraisals, much of which is directly applicable to scoring performance assessments, addressing issues such as how to train scorers. Feldman (1981) notes "people learn to attend to certain stimulus features without monitoring this process. Race, sex, cues of dress and speech, height, and so on, are all stimuli that can be automatically recorded. People may be automatically categorized via such stimuli…without intention but with future consequences for the interpretation of their behavior" (p. 129).

Bernadin and Beatty (1984) identify other obstacles that might prevent or hinder a rater from making accurate judgments. They posit that perception is "influenced by the perceiver's (i.e., rater's) experience, needs, expectations, values, dispositions, and so on" (p. 243). Furthermore they suggest that when a rater likes the person being evaluated, the rater is more likely to ignore evidence inconsistent with their overall impression, and conversely, when the rater does not like the person being evaluated, they select evidence to support their negative judgment.

Work by Ilgen and Feldman (1983) suggests minimizing the impact of irrelevant factors by giving raters specific tactics for gathering information, encouraging them to consider alternative hypotheses, and making raters aware of tendencies to pay attention to irrelevant factors.

The cognitive and social psychology literature conceptualizes three closely related dimensions of prejudice: cognitive, affective, and behavioral (Grant & Secada, 1990). A person who believes a negative stereotype about a particular group of people may feel uncomfortable in the presence of an individual from that group; however, he or she will not necessarily act in a

manner that openly displays the belief. Furthermore, a person may recognize on a cognitive level that stereotypes do not define any particular individual, but affectively, the person may still be influenced by that stereotype. An individual's belief system is closely interwoven with one's identity and understanding of the world, and is therefore extremely difficult to change.

Prejudice reduction efforts—providing strategies to develop more democratic attitudes and values towards other racial and ethnic groups (Banks, 2001; Pate, 1995)—have had some success with children and youth. Curricular interventions (such as positive descriptors of minority persons; use of multicultural materials; cooperative learning in mixed-race groups; and the use of counter-stereotypes) have been shown to increase positive racial and gender attitudes (Banks, 1995; Slavin, 1995; Freedman, Gotti, & Holtz, 1983).

The work of Greenwald and colleagues (Greenwald & Banaji, 1995; Greenwald & Farnham, 2000, Greenwald, McGhee, & Schwartz, 1998;) using the *Implicit Association Test* has shown some interesting results in identifying implicit associations between a target-concept discrimination (such as African American and European American faces) and an attribute dimension (such as pleasant and unpleasant) (Greenwald et al., 1998, p. 1465) in a manner which the researchers think is resistant to masking by self-presentation strategies. Participants are asked to associate African American faces with pleasant words and European American faces with unpleasant words (Black + pleasant) and then to reverse the associations by connecting African American faces with unpleasant words and White faces with pleasant words (White + pleasant). A participant may find the Black + pleasant combination the easier of the two if there is a stronger association between African Americans and pleasant meaning than between European Americans and pleasant meaning. As Greenwald and colleagues (1998) noted, "If the preexisting associations are opposite in direction—which might be expected for White subjects raised in a culture imbued with pervasive residues of a history of anti-Black discrimination—the subject would find White + pleasant to be easier" (p. 1465).

The societal bias exercise included in the NBPTS training uses a related technique by asking assessors to write down two or three words or phrases associated with particular bias triggers, such as, Black male, housing project. Assessors review their responses individually and then as a group discuss the possible sources of these associations. The goal of the exercise and subsequent discussion is to reveal hidden biases and thereby reduce their potential effects on the scoring process.

An important factor that affects how explicit individuals are in discussing bias is their own level of awareness of bias-related issues in society, specifically their awareness of how prejudice, bias, racism, classism, sexism, and other forms of power and privilege operate in today's social institutions (Martin, 1995). A number of researchers have examined how an individual's awareness affects the quality of their discussions—that is, the degree of explicit and forthright language employed—on bias-related issues (Helms, 1990; Khera, 1995; Szpara, 1999; van Dijk, 1984). This awareness or lack of awareness about issues of bias may also affect teachers' interactions with their students, as shown by the research cited below. Both of these research threads have impact on the present study, in that assessors are likely to exhibit varying levels of awareness of bias, thereby presenting a range of explicitness in their essays and trigger lists. In addition, assessors may carry a lack of awareness of bias into the scoring process, hence NBPTS's extensive efforts to make assessors cognizant of this.

The NBPTS Portfolio Assessor Training Manual (National Board for Professional Teaching Standards [NBPTS], 2002) provides the following text for trainers to read to assessors, summarizing the steps that the scoring process uses to help assessors actively and consciously monitor themselves during the scoring process:

> The explicit "checks" in the Scoring path are all part of a structured approach to limiting bias in assessment. They're absolutely required steps that will, in a sense, force you to be "more fair" than you might otherwise be. Working with the Note-Taking Guide, Scoring Path[2], and Personal Bias Trigger List are the keys to slowing down your evaluative process so that you can check to see if assumptions are getting in the way of seeing the underlying architecture of a performance. We know from our own experience and the experiences of hundreds of assessors that the responsibility of scoring fairly can feel overwhelming. We have created a structure that may, at times, feel cumbersome, but it is in fact necessary to give you the support you need to make the best possible evaluations. (pp. 118–119)

The necessity of holding biases at a conscious level of attention is shown in the research by Schultz, Buck, and Niesz (2000) in which they discuss data from a qualitative study of the race experiences of middle school students who attend a post-desegregated school. They suggest that,

> [In order] to educate children to participate in democratic and pluralistic culture, we need to engage students and adults who teach them in deep, and sometimes painful and conflictual, conversations about their daily school lives… it is critical to encourage students to notice and reflect on the power dynamics that shape their myriad conversations and interactions. (p. 34)

In this research, Schultz et al. discuss the outcomes of two mixed-race discussion groups. In one group, students attempted to find common ground, which the researchers described as "bridging talk," while in the other group, there was explicit disagreement across racial lines, which the researchers described as "conflict talk." The group that engaged in "bridging talk" avoided disagreement by using distancing strategies.

> Students collaborated in a denial of racism within their group by insisting on distinguishing themselves from an older generation and prevalent racist media representations. Second, they created an external Other… they focused on their similarities across racial boundaries and attempted to erase their differences by uncovering and describing their multiple and shared ethnic backgrounds. (p. 42)

Research by Pohan and Aguilar (2001) describes the use of an instrument for measuring educators' personal and professional beliefs about diversity. They discuss a significant amount of evidence that suggests that teachers' beliefs about students based on race/ethnicity, class, or gender lead to differential outcomes and treatment of students.

Numerous researchers, in fact, have investigated the impact of students' race/ethnicity, social class and gender on teachers (Guttmann & Bar-Tal, 1982; Hale-Benson, 1982; Baron, Tom, & Cooper, 1985; Brophy & Everston, 1981; Sadker, Sadker, & Long, 1993). While this research is situated in the context of teacher-student relationships, it is reasonable to consider that the same differential outcomes could be present in the assessor-candidate relationship, which involves a parallel power dynamic and evaluation arrangement.

Lewis (2001) presents analyses from a year-long ethnographic study conducted in a predominantly White elementary school in which she examined the explicit and implicit messages that teachers, students, and parents presented about race. The research was situated in a White suburb on the outskirts of an "extremely diverse metropolis" where families were mostly middle and upper-middle class, and only 2% of families were below the poverty line. The White

and predominantly middle-class teachers in this school were typical of many elementary school teachers. Lewis found that teachers and parents alike for the most part downplayed the importance of race in the school, deracialized racial incidents, or portrayed apparent color-blind philosophies. Yet, at the same time, multiple subjects within the school displayed behavior or made comments that could be construed at worst as racist and at best as being contradictory to their previous assertions that they "did not see race." Race informed who they socialized with, where they lived, where their children went to school, and attitudes they held about members of other racial/ethnic groups. Race also provided rationales for why some people were "kept down," such as their "chip on the shoulder attitude."

It is of note that, in the Lewis study, it took considerable time for many of the participants to share information about race, and on only one occasion did a participant appear to have any understanding that in fact her comments sounded racist: "Color-blindness enables all members of the community to avoid confronting the racial realities that surround them, to avoid facing their own racist presumptions and understandings, and to avoid dealing with racist events" (p. 801).

Multiple research studies have demonstrated that individuals are reluctant to openly discuss sensitive social issues such as racism and classism (Khera,1995; Lewis, 20001; Schultz, Buck, & Niesz, 2000). For that reason, the NBPTS bias exercises are consciously designed to assist assessors in examining and discussing these difficult topics.

Giroux (1998) writes that "Education works best when those experiences that shape and penetrate one's lived realities are jolted, unsettled, and made the object of critical analysis" (p. 132). Many researchers and theorists argue that people's belief systems (including their prejudices and biases) can only be affected through deep reflection and/or experiential opportunities that present information at odds with their current beliefs (Acosta-Deprez, 1995; Julian, 1996; Schön, 1987).

The NBPTS training incorporates both experience (in the various bias-awareness exercises) and reflection (through writing, discussion, and maintenance of the trigger lists), in a coordinated effort to raise awareness of bias and to empower assessors to screen out their biases when scoring.

The literature described above shows that discussing bias is a very difficult and uncomfortable process for many people, and as such, individuals utilize a range of sociolinguistic strategies to avoid or to soften discussions of bias, rather than confronting the

often taboo topic in a more direct manner (Szpara, 1999). The finding that many individuals experience difficulty in talking about issues of bias holds special import for the NBPTS's training program, since a major goal of its bias exercises is to stimulate explicit discussion of bias and its potential effects.

The literature also provides insight into the ways in which assessors' judgments can be negatively influenced by construct-irrelevant information, and offers guidance into how to mitigate these effects. This research has been incorporated into the NBPTS bias-reduction training. Based on evidence that individuals who are aware of biases at a conscious level are better able to address them in explicit form (Khera, 1995), both in conversation and action, the NBPTS's training program incorporates numerous bias-awareness exercises (described earlier) which prompt assessors to identify their biases in writing, by making trigger lists, and to refer frequently to these lists throughout the scoring process in order to maintain conscious attention to these factors.

The current study examines the effects of this training on assessors' admissions of bias and the examples of bias they provide in two written contexts: their essays and their trigger lists. Additional data from the larger research study will also permit examination of assessors' oral admissions of bias.

## Data Sources

Data were collected during June and July 2002 from five geographically diverse NBPTS training sites. The trainers were approached before the scoring session and invited to participate in the study. Trainers were selected to provide diversity in terms of self-identified race/ethnicity. Two trainers identified themselves as African American, one as Hispanic, one as White (European American), and one as Other. Four female trainers and one male trainer participated in the study. Trainers differed in their previous NBPTS experiences; one was a National Board Certified teacher (NBCT), two had served previously as assessors, and two had been trainers in previous years. The trainers were responsible for training assessors to score video entries for candidates in five different certificate areas: Middle Childhood/Generalist, Early Adolescence/English Language Arts, Early Adolescence/Mathematics, Early Adolescence/Social Studies, and Adolescence and Young Adulthood/Science. All trainers had participated in a standardized preparation program to learn how to deliver the assessor training, and specifically

how to deliver the bias-reduction exercises. Trainers worked from a training manual that provided daily plans, a timetable, talking points, handouts, and transparencies (NBPTS, 2002).

During the summer 2002 NBPTS scoring effort as a whole, there were 1,832 assessors involved in scoring portfolio entries. Of the five certificate areas from which we selected participants for this study, there were 470 assessors, approximately 230 of whom were involved in scoring video-based entries. Assessors were randomly assigned to trainers according to the regular NBPTS process. The morning of the first day of scoring, prior to the start of the training session, the assessors assigned to the selected trainer at each of the five observation sites were approached by the observer and the site scoring director (co-coordinator), and the study was explained to them. Assessors were given the option of changing scoring rooms if they did not wish to participate in the study. All assessors elected to remain in their assigned scoring rooms.

The number of assessors in each group ranged from seven to fifteen, with a total of 50 assessors participating in the study. The assessor groups varied in their racial diversity from one entirely European American group to another that included 38% teachers of color. Assessor groups also varied in the proportion of National Board Certified Teachers (NBCTs), from 0% to 46%, and in the number of years of teaching experience, from 3 years to more than 20 years. Table 1 summarizes data across the five observation sites.

Four distinct types of data were collected at each of the five sites. First, structured interviews were conducted with the trainers. The interviews were audio-taped and later transcribed. Interviews with trainers occurred immediately prior to and at the conclusion of the training. Second, the 4 day training session was audio-taped and field notes collected to support the subsequent transcription of the tapes. Nonparticipant observers, trained in ethnographic research and familiar with the NBPTS training protocol, took field notes and monitored the audio-recorders once permission was granted from the assessors in the training room. Observers were introduced as ETS staff who were observing in order to more fully understand the implementation of training protocols. The third and fourth data sources were assessors' essays and trigger lists. For the analyses presented in this report, the primary sources of data are assessors' essays and their trigger lists. Each of these is discussed in more detail in the following sections.

**Table 1**

***Summary of Observation Site Information***

| | Site 1 | Site 2 | Site 3 | Site 4 | Site 5 |
|---|---|---|---|---|---|
| Region | Southeast | Southeast | West | Midwest | Northeast |
| Trainers | | | | | |
| Training experience | 5th year | 1st year[b] | 1st year[a] | 2nd year | 1st year[a] |
| Gender | Female | Female | Female | Female | Male |
| Race/ Ethnicity | Other | African American | African American | Hispanic | White |
| Assessors | | | | | |
| Number | 8 | 13 | 7 | 15 | 7 |
| Grade levels covered by assessors | 9–12 | 6–9 | 6–9 | 3–5 | 6–9 |
| NBCTs | 1 | 6 | 0 | 0 | 0 |
| Race/ Ethnicity | 1 Asian, 1 Hispanic, 6 White | 5 African American, 8 White | 1 African American, 6 White | 15 White | 2 African American, 5 White |

[a]Previous assessor.

[b]National Board Certified Teacher (NBCT).

***Essays***

Short reflective essays (one to two paragraphs), which asked assessors to reflect on their role and their perceptions of what might hinder them in their assessment task, were collected at three points in the training and scoring process. Essay 1 was collected before the start of training on Day 1; Essay 2 was collected on Day 4, after the qualifying round (the independent scoring completed by assessors which must be of sufficient quality in order to qualify for actual scoring); and Essay 3 was collected at the conclusion of the scoring process approximately two weeks after the qualifying round.

In order to gain insights into changes in assessors' perceptions of their biases as training and scoring progressed, similar prompts were used for each of the three essays. The prompt for Essay 1 was:

As an assessor, your role will be to judge the performances of your peers. Those judgments should be as accurate and fair as possible.
Please respond to the following question—you might want to consider your definitions of teaching competence, your awareness of bias, and/or your past teaching experiences.
*Question:* What might help or hinder you in making fair and accurate judgments?

The Essay 2 and Essay 3 prompts were very similar except for the middle sentence, which was expanded to include a reference to the assessor training (in the prompt for Essay 2) and scoring experiences (in the prompt for Essay 3). The middle part of the Essay 2 prompt reads: "Please respond to the following question—you might want to consider your *assessor training*, your definitions of teaching competence, your awareness of bias, and/or your past teaching experiences." The middle part of the Essay 3 prompt reads: "Please respond to the following question—you might want to consider your assessor training, *your scoring experiences*, your definitions of teaching competence, your awareness of bias, and/or your past teaching experiences.

Despite these slight differences in the wording of the essay prompts, the question to the assessors of what might help or hinder them in making fair and accurate judgments remained the same across all three prompts. The collection of assessors' responses to the essay prompts was the only aspect of their training experience that differed from the general assessor training protocol.

Four days of training elapsed between the assessors completing Essay 1 (on a Monday morning) and Essay 2 (on Thursday afternoon). During that time, assessors were trained to score a particular entry, experienced the various bias-reduction exercises, and had opportunities to practice scoring 12 to 14 candidate responses. The Friday of the first week was the first day of scoring actual candidate performances, and scoring continued for the next two weeks. Essay 3 therefore was completed 15 days after Essay 2. Forty assessors completed all three essays, nine completed two essays, and one assessor completed only the first essay. The one assessor who wrote only the first essay withdrew from the training for personal reasons.

In the nine instances in which assessors only completed two essays, it was not possible to determine whether those assessors did not complete the scoring process and were not present on

the final day of scoring, or whether because of other time commitments they did not complete the final essay after their scoring obligations were finished. The incomplete third essays all came from the two largest groups of assessors who had been observed (four out of 15 from Site 2 and five out of 13 from Site 4). Given the high rate of successful completion of the assessor training (greater than 95% success rate), it is unlikely that approximately one third of assessors at these two sites would not have successfully completed the qualifying round (DeLuca, 2003, personal communication). It is more likely that individuals did not take time at the end of their scoring session to complete and submit the final essay, or that the trainer was occupied with other matters and unavailable to remind the assessors, rather than any objection to the data collection itself. Even though 20% of the assessors in the study did not complete all three essays, it was still possible to investigate changes in assessors' perceptions of bias since all but one assessor completed essays both before and after the training was completed.

### *Trigger Lists*

The second data source for the study was the assessors' trigger lists. The trigger list is a sheet of paper that is given to assessors on the first day of training. It is oriented in landscape format and folded in half. The sheet is printed on blue paper, along with several other critical pieces of the scoring apparatus (rubric and note-taking guide) as another means of emphasizing the importance of assessors keeping the trigger list in front of themselves. Figure 1 shows the layout of the inside of the trigger list.

| TRIGGERS THAT MAKE ME SCORE HIGHER | TRIGGERS THAT MAKE ME SCORE LOWER |
|---|---|
| | |

***Figure 1*. Layout of the assessor trigger list.**

As shown in Figure 1, on the left-hand side the text at the top says, "triggers that make me score higher" and on the right-hand side the text says, "triggers that make me score lower." Assessors tend to keep the sheet of paper folded in half so that someone walking around the room is unable to read what is written on it. Assessors are directed to look at this list at the start of scoring each candidate response as part of the scoring process. At the end of the 4 days of training, observers at four of the five sites made a request to the assessors that they share their trigger lists so that a copy could be made for research purposes and then returned to them. Assessors were under no obligation to do so, but were assured that no identifying information would be collected.

The observer at Site 1 did not ask the assessors to share their trigger lists, due to concern expressed by the site coordinators that it would interfere with the scoring process. At Site 2, eleven of the thirteen assessors shared their lists; at Site 3, four of the seven assessors; at Site 4, nine of fifteen of the assessors; and at Site 5, all seven assessors shared their trigger lists. Of the 42 assessors to whom the request was made, 31 (74%) shared their trigger lists. The trigger lists were collected without any identifying information, as the researchers did not intend to match the assessors' essays to their trigger lists. Therefore, it is not known if some or all of the assessors who did not complete the third essay also did not share their trigger lists.

## Methodology

To support the investigation of assessors' awareness of bias before, during, and after the NBPTS training, it was necessary for the researchers to develop detailed coding protocols to guide analysis of the two data sources used in this study: assessors' essays (written in response to the three prompts described earlier), and their individualized trigger lists. The coding approaches used for these two data sources are described in depth in this section.

### *Coding the Essays*

Rather than impose a formal classification system on the data, inductive data analysis methods (Bogdan & Biklen, 1992; Erickson, 1986) were used to review the essays in order to construct, test, and refine interpretative themes (presented below). Specifically, the researchers used consistent indexing and retrieval of information to develop categories structured on common agreement among concepts in each category (Michalski, 1993; Thomas, 2003).

Language content analysis (van Dijk, 1987) and interactional sociolinguistics (Schiffrin, 1994) were applied to examine the manner in which assessors identified potential helps or hindrances to making fair and accurate judgments. Language content analysis or text analysis provides a systematic approach to examining open-ended textual responses such as those given by the assessors in their essays. The texts were "mined" for lexical items and syntax structures which clustered into predefined categories or which represented new categorizations in response to the research questions. Using language content analysis, researchers examined the essays for patterns of bias awareness and specific evidence of increasing awareness over time. These included the degree of explicitness in language or terminology, the extent of the usage of terms introduced during training, and the personal application of the bias information provided in training.

Interactional sociolinguistics views discourse as a social interaction in specific contexts (Gumperz 1999). In other words, the assessors' responses to the essay questions (as well as their recorded trigger lists) were analyzed within the framework of a training program in which they were being evaluated, a context in which their ability to perform to certain expectations would determine their ability to remain in the portfolio scoring program.

Under the assumption that discourse is socially embedded, language can be considered as a tool used to achieve certain ends or means (Hymes, 1962). Assessors might use language to demonstrate their increasing awareness of biases, or to make a claim of professional skills in evaluation, years of experience in teaching, or lack of biases. While language can be manipulated purposefully by a speaker, language can also be revealing, in that speakers make unconscious choices about minute linguistic details of their discourse. For example, a choice of *I* versus *you* versus *we* when discussing an undesirable task that must be completed can convey meanings of superiority and dominance (*I* or *you*), or it can convey solidarity and a sense of community (*we*).

The assignment of evaluative words to a speaker's pronoun choices is subjective and must be done in the context of a deep understanding of the social context in which the particular interaction occurs. In the present study, the researchers observed all 4 days of the training at each site, interviewed the trainers both formally and informally, and spoke informally with the assessors during break-times and meals, in addition to collecting the audio-taped and written data. This allowed for a richer understanding of the context of the training and of the assessors' discourse within that social environment.

Specific language patterns that indicate increasing awareness of bias include use of first person pronouns, such as *I* and *my*, versus second- or third-person pronouns, such as *you*, *your*, or *one's*. The use of active versus passive verbs and adverbs for emphasis was also examined. Active verbs were used to indicate ownership of bias, whereas passive verbs were used to provide distance between the assessor and the bias(es) being discussed. In some cases, the assessor was not present in the sentence at all; biases were referred to wholly in the abstract, with no connection to personal ownership. Adverbs such *truly* and *really* were used in some cases to emphasize the absence of bias, and in other cases adverbs such as *extremely* or *very* were used to emphasize the degree to which the assessor was now aware of their biases and able to reduce their impact through deliberate action. Finally, the use of synonyms for bias, such as "feelings" and "reactions," were considered. The use of synonyms allows assessors to distance themselves from the socially taboo term of bias.

Evaluation of sentence structure in the essays included many of the linguistic aspects described above, taken as whole. Brief declarative sentences stating ownership of specific biases were rare; rambling sentences with many interim clauses were more common, allowing the assessor to distance himself or herself from a sensitive topic. These specific linguistic techniques were considered assuming that the majority of assessors might want to unconsciously distance themselves from topics that are largely left undiscussed among European American, middle-class members of U.S. society. Reflective of the current pool of teachers in the U.S., the majority of assessors in this study were European American. While socioeconomic data were not available for the assessors in this study, all were, by definition, currently employed school teachers, who may be considered as part of the middle class in U.S. society.

In addition to the language analysis that was conducted, each essay was separately reviewed and coded by two coders using the schema shown in Figure 2. One coder analyzed the full spectrum of essays across the five sites to determine the major categories outlined. These categories were then reviewed and refined jointly with a second coder. Each coder coded a practice set of essays independently, and then they compared results. Through repetition of this process, the two coders were able to consistently identify text written by the assessors that fit each of the categories. During the coding process for the remaining essays, exemplars were shared and analyzed to ensure that the coders continued to maintain agreement on how to apply the various categories.

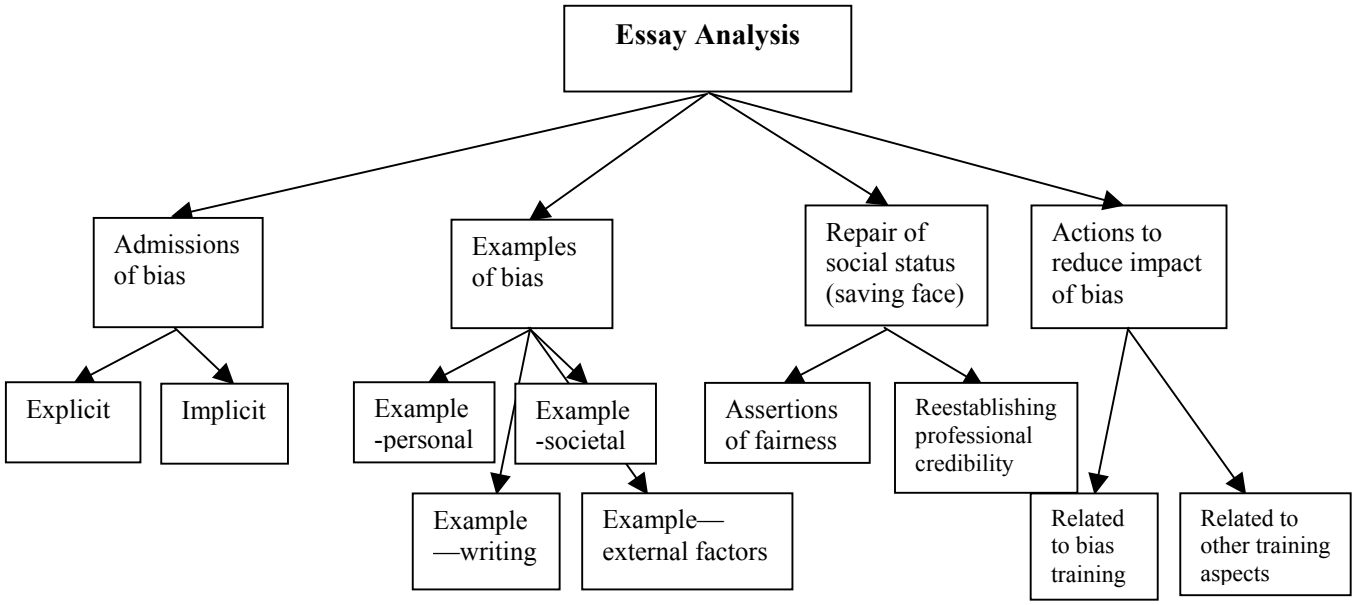


***Figure 2.*** **Outline of coding approach for the essays.**

Two distinct but interrelated coding categories were initially developed: an admission of bias and an example of bias. Linguistic techniques used by assessors to address the socially taboo issues of bias, or to avoid directly discussing issues of bias, were also coded, as were specific behaviors that assessors identified as utilizing to help them make fair and accurate scores. These coding categories are discussed in greater depth below.

The first coding category focused on patterns of bias admission and specific evidence of increasing awareness in the assessors' essays. A bias admission was considered to be a statement made by an assessor which indicated (self-)awareness of one or more specific factors or triggers that could limit or assist him or her in assigning fair and accurate scores. Although the degree of explicitness in admitting bias exists on a continuum, for the sake of the present analysis, the researchers divided the continuum into two categories. Thus, assessors' admissions of bias were coded as being either an explicit admission or implicit admission of bias.

The second coding category was an example of a bias, which was considered to be an illustration or a description of a bias. In instances in which assessors were explicit about their admission or ownership of bias, they often supplied particular examples. Alternatively, an

assessor might identify examples of bias without claiming to have the bias(es) described (admission).

Examples of bias were coded using the same categories that are used in the NBPTS training: personal biases, societal biases, and writing biases. External factors were noted in a separate category. External factors included those outside of the general definition of bias, and usually involved physical factors such as the assessors being tired, hungry, cold, etc., that might affect his or her ability to concentrate.

The third coding category dealt with instances in which assessors made assertions of fairness or used linguistic techniques to save face. These strategies were seen as belonging to a larger category called "repair of social status."

The fourth and final set of codes applied to assessors' essays marked instances in which they suggested actions that they could take to reduce the impact of bias. This category was divided into two: actions that were specifically tied to the bias training (such as referring to the trigger lists), and actions related to other aspects of the training (such as referring to the benchmark cases).

Triangulation of the data was accomplished through a variety of techniques at both the data collection level and the data analysis level. In the former, collection of data at five different sites, with five different trainers, allowed for comparisons both within and across sites. In addition, the collection of both essays and trigger lists allowed for two different sources of information to provide insight into the bias training process. As noted previously, although not all assessors submitted Essay 3, sufficient data were available for comparison across sites. In analyzing the data, triangulation was provided by using two coders.

In later analyses, audiotapes of the training session will provide further opportunities for triangulation of data.

### *Coding the Trigger Lists*

The trigger lists were used as a means of providing a different perspective on assessors' thinking and understanding of the issue of bias. The training provides multiple opportunities—at the end of each bias-awareness exercise and after reviewing benchmarks and training cases—for trainers to direct assessors' attention back to their trigger lists to update their lists with new information that they might have learned about their own personal triggers and biases.

In order to look for patterns across trigger lists, the triggers were classified into 19 categories drawn directly from content-based themes in the data. The initial categories were developed after multiple reviews of the trigger lists, in which reviewers grouped triggers according to similarities in content. As the initial categories began to form, they were tested through application to additional lists of triggers. For those triggers that did not fit, the categories were either reorganized or new categories were added to best reflect the data. For example, two small categories (teaching and content) were eventually combined into a larger group called "fatal flaws," which represented the data points within the group better than the earlier two categories did. ("Fatal flaws" are critical faults that an assessor may use to discredit other parts of the response.)

As a result of this process, the 19 data-driven categories were grouped into four larger categories (personal biases, societal biases, writing biases, and fatal flaws), following the NBPTS training manual's classification of bias types. Table 2 provides examples of triggers listed by assessors in each of the four major categories.

After the comprehensive set of categories was developed and tested against the complete set of data, one researcher and two NBPTS trainers independently categorized the words from the assessors' trigger lists. Agreement between any two pairs of categorizers ranged from 73% to 79%, with unanimous agreement on 69% of the triggers. Discussions with the trainers led to the resolution of the majority (approximately 95%) of the differences, and the remaining triggers were classified after discussion with an additional categorizer and by looking at the context of particular triggers.

In some instances, contextual information such as the position of a particular trigger in the written list was vital in making a classification decision. For example, some assessors used "disorganization" in reference to the written commentary, such as one who listed triggers of "disorganized and confusing writing; grammar and spelling errors." In such cases it was clear that the term "disorganized" referred to writing biases, because it appeared in the middle of a list of other words that were clearly identifiable as being writing triggers. In other instances, the context was different. For example, one assessor began the list of triggers with "disorganization; lack of confidence; over-ambition …" Here, the placement of "disorganization" early in the trigger list makes it likely that the term refers to a personal characteristic rather than a writing style, and thus it would be coded accordingly.

**Table 2**

***Examples from the Trigger Lists***

| Bias groupings | Examples of assessors' triggers |
|---|---|
| Personal bias | Neat and professional appearance |
| | Enthusiastic/smiling/body language |
| | Pretty bulletins |
| | Bare classroom—poor appearance |
| Societal bias | Type of school (urban/suburban/bigger) |
| | Socio-economic underprivileged students |
| | Less money spent per student |
| Writing bias | Organized, concise writing |
| | Flowery writing |
| | Documenting references |
| | Disorganized, scattered information |
| Fatal flaws | Goals not appropriate for class |
| | Expectations for gifted students too low |
| | Not including all students in class discussion |
| | Poor safety habits |

## Results

Given the nature of the essays (in which assessors responded to general, open-ended prompts) and the trigger lists (specific lists generated during training by assessors in response to learning about their individual triggers), we expected to learn different information about assessors from these two sources of evidence. We also anticipated, however, that the data from one source would complement and inform data from the other.

This section is divided into three parts: the first presents the analysis of the assessors' essays, the second presents the analysis of the assessors' trigger lists, and the final part compares the results from the two data sources.

### *Essays*

As described earlier, assessors responded to the essay prompts prior to the start of training, at the end of the 4 day training session, and finally at the end of the scoring session 2 weeks after training was completed. The completion of the essays by the assessors was an addition to the

regular training protocol. The following sections describe the results of the analysis of assessors' essays with respect to admissions of bias, examples of bias, assertions of fairness or status, and notation of actions to reduce bias.

*Admissions of bias.* There was considerable variation among assessors regarding the degree to which they showed evidence of ownership of their biases. Some assessors wrote statements that contained no references whatsoever to bias, and some wrote implicit admissions of bias. The majority were much more explicit in admitting their biases. This section begins by illustrating the ways in which assessors admitted in any of their essays to having biases, and then focuses on how those admissions changed over the course of the training and scoring.

The following series of quotations from assessors' essays illustrates the ways in which assessors admitted to having biases, from no admission of bias, to implicit and explicit admissions of bias. In every instance that follows, where an assessor's essay is quoted, the excerpt is identified by the essay it came from (Essay 1, Essay 2, or Essay 3), the site, and an assessor identification number. Italics are sometimes used to draw the reader's attention to certain features in the assessor's response that are discussed in the text.

In the first example below, the assessor states she cannot identify anything that might get in the way of fair and accurate judgments. This was a common technique used by assessors in answering the question of what might help or hinder them: they reassured the reader that nothing would hinder them; they were already fair in their judging abilities. The writer further emphasizes this reassurance by adding the adverb "truly," but then allows for the possibility that she is not correct, by using "I feel" as opposed to "I know," and by following her assertion with the clarification that she "can't think of anything." Both of these linguistic techniques allow the assessor to later change her stance without loss of face.

> I truly feel that *there was little to hinder me* [emphasis added] in making fair and accurate judgments. If anything, actually I can't think of anything.
> Assessor 4, Site 1, Essay 3

The next two examples represent the ways in which assessors made implicit admissions of bias without taking personal ownership. In the first example, by using the second person singular "you," the assessor distances herself from the bias. In the second example, the assessor

uses a slightly different approach. She does admit group biases and owns them by referring to "our" biases, but places the burden of responsibility on the trainer with the phrase "she surely clarified our biases."

> I think the bias training is invaluable as an assessor. It makes you realize your personal "triggers" that you aren't even aware you have.
> Assessor 4, Site 4, Essay 3

> She [the trainer] surely *clarified our biases* [emphasis added] and presented the various pictures of teaching competence.
> Assessor 1, Site 1, Essay 3

The next three examples represent the ways in which assessors made explicit admissions of bias, taking personal ownership. The first example presents two excerpts from the same assessor, one from Essay 2 and the other from Essay 3. In the first excerpt she clearly identifies the biases and emotions as her own; in the second excerpt she is completely open about the idea that, on occasion, bias did impact scoring judgments. She emphasizes the normality of this occurring by using the interjection "of course" when admitting that biases affected her scoring. Her writing in that essay continues to clarify that, on those occasions when she felt that the bias was significant, she "bounced ideas off the trainer." In the second example the assessor also makes an explicit admission of personal bias. In the third example, while the assessor does not say "my biases," reference is made to "things that would make me score higher or lower," a standard phrase used in the NBPTS assessor training to describe biases. Ownership of these biases is admitted by the use of the personal pronoun "me" in two places in the sentence.

> What helps is being aware of my biases and emotions while scoring (Essay 2) … *Biases, of course, hindered* [emphasis added] my judgment (Essay 3)
> Assessor 5, Site 1, Essay 2 & 3

> Sometimes I have the issue of *personal bias* [emphasis added] which I have to check my trigger list to monitor myself.
> Assessor 1, Site 3, Essay 2

> The bias training helped me to make fair and accurate judgments. Being aware of things that would make me score higher or lower was highly significant *for me* [emphasis added] to be fair and accurate.
> Assessor 1, Site 5, Essay 3

Given that an overt focus of the NBPTS training is to educate assessors about the influence of bias in scoring and the need to actively identify one's own biases and to maintain a conscious level of awareness of "things that would make me score higher or lower" (a standard phrase used in the NBPTS training), it would be reasonable to expect to see an effect of this training in assessors' discussions of what helped and/or hindered them in making fair and accurate judgments.

Table 3 demonstrates that by the end of the scoring process the majority of assessors had made explicit admissions of bias in their essays. The column on the left identifies how the assessors were categorized based on Essay 1 (before training), having made either no admission of bias, an implicit admission, or an explicit admission of having any biases that might interfere with the scoring process. Twenty-one (42%) of the 50 assessors responded to Essay 1 without making any such admission, explicitly or implicitly. Given that the majority of assessors were assessing for the first time, and so had not been through bias training in previous years, this is not wholly unexpected.

The column on the right in Table 3 identifies the type of admission made across Essays 2 and 3 (after training), for example, showing that of the 6 assessors who initially made only an implicit admission of bias in Essay 1, five made explicit admissions in Essay 2 or Essay 3. As Table 3 shows, the majority of the assessors who had made either no admission or an implicit admission of bias in Essay 1 increased in their level of awareness of bias, and in their personal ownership of bias, after training. One assessor only completed one essay, but of the remaining 49 who completed two or three essays, 43 (88%) assessors made an explicit admission of bias at some point in at least one of the essays. Five assessors only made at most implicit admissions of bias across all essays, and one assessor wrote all three essays without any recognition of bias.

**Table 3**

***Admissions of Bias Before and After Training***

| Essay 1 | Essay 2 or 3 | *N* |
|---|---|---|
| No admission of bias (*N* = 21) | No admission | 1 |
| | Implicit admission only | 4 |
| | Explicit admission | 15 |
| Implicit admission of bias (*N* = 6) | No admission | 0 |
| | Implicit admission only | 1 |
| | Explicit admission | 5 |
| Explicit admission of bias (*N* = 23) | No admission | 0 |
| | Implicit admission only | 1 |
| | Explicit admission | 22 |

*Note.* One of the 21 assessors who made no admission of bias only completed one essay.

While the majority of assessors made explicit admissions of bias in their essays after training, it is instructive to look at several examples where this was not the case. Discrepant cases such as these allow for richer understanding of the data set as a whole. The one assessor (Assessor 2, Site 1) who completed all three essays without ever making even an implicit admission of bias discussed in her Essay 1 contributing factors to fair scoring such as explicit standards, practice opportunities, rubrics, and feedback from the trainer. The only hindrances that she could envisage were external to her—such as being rushed or pressured to score in too short a time span. In the second essay she mentions that "the bias training was very helpful," as were example cases. A similar statement was made just over two weeks later when she completed Essay 3 and again stated that the "bias training helped tremendously," but there was no elaboration on how or why that might have been the case.

Assessor 11 at Site 4 was one of the small group of assessors who initially made no admission of bias and who made only an implicit admission in the second essay. There is no third essay for this assessor. In the first essay, she identified herself as being open-minded and willing to "allow my trainer to guide and teach me on ways to view the material through NBPTS eyes." She merely writes a question mark after the word "hinder." On Essay 2 she makes reference to bias but uses the third person to remove herself from the bias, and thus not claiming personal ownership of the biases mentioned:

> As pointed out by our trainer, biases sometimes appear. Assessors must recognize there is a bias and remove *themselves* [emphasis added] from it.
> Assessor 11, Site 4, Essay 2

Assessor 13 (Site 2) was another assessor who did not make as much progress as some others in terms of admitting bias. In the first essay she makes reference to the importance of examples in training, and this thought is reiterated in the second essay without any direct or indirect comments about the role of bias. In the third essay, the assessor writes that:

> The trigger list helped me in making fair judgments. During the training we were to write down things that we felt would bias our judgment. Referring back to this was very beneficial. I don't feel that anything hindered me in making fair and accurate judgments. I just referred back to the trigger list.
> Assessor 13, Site 2, Essay 3

The notion of change in assessors' attitudes towards the recognition of the existence of bias is further supported by comments made by a number of assessors in their essays. Ten percent ($N = 5$) of assessors made specific reference to uncovering biases that they had never known about previously, as illustrated in the two examples below. In both examples the assessors did not specify what their "unconscious" or previously un-"realized" biases were.

> The initial training was intensive, but extremely beneficial. It helped me put insight into my known and *unconscious* [emphasis added] biases.
> Assessor 5, Site 5, Essay 3

> I discovered today when I got to the qualifying round 2 *biases I hadn't realized I had* [emphasis added].
> Assessor 6, Site 5, Essay 2

In summary, while assessors varied in the degree to which they admitted bias, there was a substantial change over time in the level of bias admission made by assessors, from 42% explicitly

admitting bias in Essay 1 to 86% in Essay 2 and Essay 3. These data are supported by comments from assessors who mentioned that they uncovered unknown biases during training. Clearly, the training has an impact on assessors' understanding of their personal ownership of bias.

*Examples of bias.* Not only did most of the assessors take ownership of bias (by making admissions of bias) as described in the previous section; many also provided statements illustrating particular types of bias. Based on research on general awareness of bias (Delpit, 1995; Derman-Sparks & Phillips, 1997) and on individuals' reluctance to explicitly discuss issues of bias (Helms, 1990; Khera, 1995; Szpara, 1999; van Dijk, 1984), it was expected that the number of examples of bias would be greatest for personal bias, followed by writing bias, and lastly societal bias. This section begins with a general discussion of the types of examples that assessors supplied, and then briefly looks at how those examples changed over the course of the essays.

As will be illustrated in this section, assessors sometimes used terminology used during the NBPTS training to refer to bias, specifically personal bias and writing bias. Nine assessors used phrases that came directly from the training, referring to "personal bias," "personal triggers," or "personal likes and dislikes." Only one assessor explicitly used the phrase "writing bias," while another identified "writing techniques" as a bias. There were no instances of an assessor using the term "societal bias."

In the majority of instances, however, assessors did not specifically label their biases as personal, writing, or societal biases. When the examples supplied were not specifically labeled, the researchers used the definitions used in the NBPTS training to categorize the examples as personal, writing, and/or societal biases. If the assessors provided an example of bias such as volunteering information that would "hinder them from making fair and accurate scores", it was coded accordingly, whether or not the assessor actively labeled it as a bias.

The first quote from assessors' essays below makes reference to personal bias, while the second relates to writing bias.

> Being aware of my *personal bias* [emphasis added] helps me see clearly instead of viewing a performance through my preferences.
> Assessor 3, Site 2, Essay 2

> I still have to be aware of *my writing bias* [emphasis added]: - grammar—sp.
> Assessor 2, Site 4, Essay 2

The following five excerpts from assessors' essays illustrate how assessors provided examples of the types of bias that they might encounter. In the first quote, the assessor refers to "personal reactions" in lieu of "personal bias" or "personal triggers." Specific areas of possible bias are mentioned, such as classroom management style, speech, and student-teacher interactions. However, the assessor does not indicate explicitly what direction those biases might take, that is, what aspect of classroom management style might make the assessor score higher or lower.

> Also I will have to constantly monitor myself to make sure I do not assess based on my personal reactions to *classroom management style, speech, student-teacher interactions, etc.* [emphasis added].
> Assessor 6, Site 2, Essay 1, coded as specific examples of personal bias

In the following two instances, assessors recognized the potential impact (both negative and positive) that a candidate's writing could have on them as assessors.

> As a teacher I scrutinize students [sic] work very closely to help them improve, and to be reminded that my task is to assess a candidate's professional performance *rather than analyze sentence structure or speech/grammar patterns* [emphasis added] allowed me to stay focused.
> Assessor 9, Site 2, Essay 1

> It was good to make us aware of things that might make us score lower (or higher) such as how well (or poorly) the written commentary was written with respect to *grammar, spelling, etc* [emphasis added].
> Assessor 4, Site 5, Essay 3

As previously noted, no assessor used the phrase "societal bias," and this category had the lowest number of examples that were coded as such. Two examples are provided below. In the first example, the assessor indicates that inappropriate or dated labels on populations or cultures would be a hindrance, indicating that a bias on the part of the candidate could cause bias on the part of the assessor. In the second example, the assessor provides specific details about aspects of writing that are triggers for him or her. The assessor also demonstrates awareness of the potential impact of societal bias by listing issues of ethnicity/race, urbanicity, and gender, although claiming that none of these areas has an impact on his or her judgment.

> Hindrance—*inappropriate or dated labels on populations or cultures* [emphasis added] i.e. Afro American in year 2002 not political [sic] correct or insightful to identity of a culture.
> Assessor 1, Site 2, Essay 2

> I never thought font, spacing, or technical aspects of writing would make a difference, but it does. I am not as aware of *color, urban/rural, or sex of teacher* [emphasis added].
> Assessor 8, Site 4, Essay 2

Assessors occasionally employed indirect ways of responding to the essay prompts, utilizing euphemisms to describe bias while still providing examples of what might hinder them from rendering fair and accurate judgments. In the two excerpts provided below, bias is referred to as "something that will draw [an assessor's] attention" or as something that "influences feelings." Rather than actively claiming writing as a bias, these assessors provide the example but soften their connection to it. In both cases, the assessor makes the bias the "actor" in the sentence, drawing attention or influencing feelings. This allows these assessors to be the passive recipients of the action, thereby reducing responsibility on their part.

> I think a paper that contains a huge amount of grammar mistakes will be *something that will draw my attention* [emphasis added].
> Assessor 2, Site 4, Essay 1

I also know that grammar *influences my "feelings"* [emphasis added].

Assessor 6, Site 4, Essay 1

The data were analyzed to investigate whether the assessors provided more examples of bias in the later essays. Given that there were fewer specific examples of bias than admissions of bias, no strong pattern emerges from the results. Table 4 summarizes the data. The proportion of assessors who provided an example of bias in the first essay (41%) was slightly lower than the percentage who made an explicit admission of bias in the first essay (46%).

Further, the number of assessors who provided examples of bias did not change as dramatically after training. Of the 49 assessors who responded to more than one essay, 14 (29%) did not provide an example of bias in any of their essays. As Table 4 illustrates, initially in Essay 1, 29 assessors did not provide an example of a type of bias that might hinder them in the scoring process, but 15 of them provided an example in at least one of the essays completed after training.

**Table 4**

***Examples of Bias in the Essays Before and After Training***

| Essay 1 | Essay 2 or 3 | *N* |
|---|---|---|
| No example (*N* = 29) | No example | 14 |
| | Provided an example | 15 |
| Provided an example (*N* = 20) | No example | 5 |
| | Provided an example | 15 |

In the 139 essays written by the 49 assessors, 78 (56%) essays did not contain any specific examples of bias. Across the remaining 61 essays, 77 triggers were supplied that could be categorized as personal, writing, or societal bias triggers: personal bias examples were supplied 49 times (64%); societal biases, six times (8%); and specific writing biases, 22 times (29%). As anticipated, assessors were most likely to provide a bias example that could be categorized as personal bias, and least likely to provide an example of a societal bias.

It is instructive to see how assessors changed the ways in which they discussed the issue of bias over the series of essays. Several examples of comments from the same assessor over

time are supplied below to illustrate that there was some degree of progression in explicitness and/or number of biases mentioned. This degree of progression represents achievement of one of the National Board's goals for the assessor training: to raise awareness of biases. During one of the bias-reduction exercises, trainers remind assessors of the goals of the exercise:

> The goal of our work is *not* to prove that you have no biases. This is an exercise with yourself to find out what triggers your biases and what direction your biases run in. To the degree you succeed in identifying your biases, you can control them during scoring, and thus become a better assessor.
> (NBPTS, 2002, p. 19)

It is possible that some assessors increased their awareness of biases they held but did not share this in their essays. However, since the essay prompts specifically addressed awareness of bias, the number of assessors who did not disclose bias awareness should be small.

In the first example below, from Essay 1, the assessor identifies language usage as a potential trigger, but uses the passive voice to discuss it. There is no personal pronoun or other explicit connection to the assessor, and the main verb is conditional, "might be." In the second example, after 4 days of training, the assessor uses an active sentence construction, the personal pronoun "I," and a clear connection to bias: "I have a bias…." In addition, the assessor uses more specific language in describing the bias, from "incorrect usage of language" to "grammatical and spelling errors."

> One hindrance might be the *incorrect usage of language* [emphasis added] in the writing component the candidate might use.
> Assessor 4, Site 4, Essay 1

> While going through training I realized I have a bias in looking for grammatical and spelling errors. It is a trigger I must be careful to watch for.
> Assessor 4, Site 4, Essay 2

Another way in which assessors' descriptions of particular biases changed over time is illustrated in the pair of excerpts below:

> What helps me to make fair and accurate judgments is the ability to step back and see the whole picture rather than focus on the minute. I don't see female/male, I see teacher/facilitator; *subject matter is irrelevant, how that concept is presented is* [sic]*, and so forth* [emphasis added].
> Assessor 3, Site 1, Essay 2

> The only thing I had to be especially careful about was *letting the type of class (learning level or subject matter) affect my attitude* [emphasis added].
> Assessor 3, Site 1, Essay 3

In this set of examples, the assessor clearly deepened her understanding of bias and changed from denying the impact of subject matter to recognizing the impact that it could have on her attitude—and therefore on scores that she might award. Given that this change occurred from Essay 2 to Essay 3, this deepening of awareness occurred after the training, at some point during the 2 week scoring window. It would be interesting to know whether the trainer conferenced with this assessor after having completed an adjudication (or score deliberation) involving this particular form of personal bias.

The final example provided below reveals how an assessor changed over the duration of the training and scoring period. An excerpt from each of the three essays is provided. This assessor changed from not being able to identify any potential biases in Essay 1 to being able to supply a very specific example regarding writing. The 2 week scoring period seemed to reinforce the irritation expressed in Essay 2 about this particular style of presenting information in the written commentary since the level of irritation appears to have increased in Essay 3. It is important to note, however, that this assessor recognized the particular details of what bothered her, and also in Essay 3 provides an example of specific action that she took in the light of finding this type of writing.

> At this point I am really not sure what might hinder me from making a fair judgment.
> Assessor 10, Site 2, Essay 1

> I will be helped in making a fair judgment because of the bias training that was given. I may be hindered in giving a fair judgment when the candidate constantly tells the assessor what standard to score their information under. This is annoying.
> Assessor 10, Site 2, Essay 2

> What hindered making a fair judgment were candidates who directly stated standards and/or called themselves an accomplished teacher. This arrogant writing was extremely annoying. Several times I had to reread passages to ensure I was accurately assessing.
> Assessor 10, Site 2, Essay 3

In summary, assessors did not supply specific examples of bias as frequently as they made admissions of bias. Personal biases represented the majority of the biases for which assessors supplied specific examples, followed by writing biases. It is striking to observe the paucity of references to biases under the general umbrella of "societal biases." This may be due, in part, to the sensitive nature of discussing societal bias. In addition, preliminary reviews of the field notes taken during the 4 day training at each site indicate that a similar amount of time is formally dedicated to personal, societal, and writing bias-reduction training (approximately one and a half hours each). However, in terms of additional, informal time devoted to discussions of bias—that is, time spent on discussion of biases interwoven into other aspects of training (such as the review of benchmarks and training cases)—there was significantly more time spent on personal and writing biases that assessors uncovered as they reviewed additional cases. Similar discussions rarely, if ever, happened for societal biases.

Assessors did not change as dramatically in how often they provided examples of bias as a result of the training. Just under one third of the assessors never provided examples of bias. Among those assessors who provided examples, it was possible to identify particular assessors

who deepened their own understanding of the types of bias that were triggers for them during scoring, through increasing specificity of those examples.

*Repair of social status.* Various linguistic techniques for distancing oneself from bias have been discussed. These include the use of second or third person pronouns or absence of any personal reference in the sentence; use of adverbs to emphasize surprise in finding bias or to assure the reader of the assessor's ability to score fairly; and the use of indirect sentence structure to create distance between the actor/writer of the sentence and the admission of bias. Another linguistic technique to add to this list is using assertions of fairness to save face after having admitted bias. While most of the other techniques involve specific linguistic strategies at the grammatical or syntactic level, asserting fairness to save face is different in that it involves a topic change rather than manipulations in grammar or syntax.

In this section, we discuss two distinct but related patterns found in assessors' essays: assertions of fairness, and claims of professional status or ability. In both instances, the goal appeared to be an attempt to repair the assessors' social status, or to "save face."

Linguistic "loss of face" was first proposed by Brown and Levinson (1987), and was expanded upon by Gumperz (1982, 1999), Scollon and Scollon (2001), and others. According to Scollon and Scollon, "Face is the negotiated public image, mutually granted each other by participants in a communicative event" (p. 45). If the larger communicative event involves the assessors' participating in training and evaluation by the NBPTS for the purpose and privilege of scoring NBPTS portfolios, then the assessors may regard the essays they wrote as part of their training process—a place where they were expected to admit biases (as the training encourages them to do)—also as a place where they must maintain their best "face" or image as a fair assessor.

All but one assessor made at least one implicit or explicit acknowledgement that biases could hinder their scoring process, and 44% of essays contained at least one specific example of bias. A number of assessors, however, made assertions of fairness that counter-balanced their admissions of bias.

Part of the NBPTS assessor training deals with the issue of familiarity, and the possibility of assessors being overly severe (or generous) if they were familiar with the particular content presented in a candidate's response. Specifically, the training instructs assessors to set aside their own teaching experiences and to rely on the NBPTS's definition of teaching quality for their

judgments. Assessors are also directed to recognize that "open-mindedness" does not exist in a vacuum; all assessors hold biases, and so open-mindedness is not a reliable factor to assure fairness in scoring.

Nonetheless, 36% of assessors claimed that they would be fair because of their teaching experiences (grade levels, subjects, or schools), 30% of assessors claimed that they were fair because of their open-mindedness or objectivity, and 12% felt that the process itself was fair because of the fact that candidates' submissions were anonymous. Overall, 62% of assessors expressed one or more of these reasons for considering themselves likely to be fair assessors. Such assertions were often accompanied by comments that indicated an inability to think of anything that would hinder them from making fair and accurate judgments. The majority of these comments occurred in Essay 1, decreasing in Essay 2 and Essay 3 as explicit admissions of bias increased in those essays.

The first set of excerpts below is comments made by assessors justifying their fairness through claims of personal beliefs and attitudes.

> I do know what will help me, I have a *personal belief in fairness* [emphasis added] throughout education and life in general. I believe for this process everything that is to be evaluated must be done fair. If it is not, then the process becomes severely flawed.
> Assessor 10, Site 2, Essay 1

> The fact that *I am* personally I feel, *objective* [emphasis added] and follow… rubrics given well. This will also lead to an unbiased assessment.
> Assessor 7, Site 1, Essay 1

> Also I am *extremely conscientious* [emphasis added] about giving accurate, fair scores to my student and I'm sure that attitude will serve me well as an assessor.
> Assessor 6, Site 2, Essay 1

The following excerpts are comments made by assessors who make a variety of claims about fairness, tying in their own content knowledge and a range of school experiences in terms of grade levels, subjects, or school location.

> The things that will help me be a fair and accurate assessor would be *knowledge of my field* [emphasis added].
> Assessor 4, Site 2, Essay 1

> For the past three years I have been *teaching in several different environments. I have taught various age levels and subjects,* [emphasis added] which required me to be versatile in my teaching practice…given me insight to make fair and accurate judgments.
> Assessor 8, Site 2, Essay 1

> I have *taught in both urban and rural districts*. So I understand that there can be many ways to reach different types of students. It has caused me to be *very open-minded* [emphasis added].
> Assessor 9, Site 4, Essay 3

Although assessors occasionally recognized that their teaching experiences could in fact create opportunities for bias, these comments were uncommon.

> I find that if I have taught a unit I am more critical … also I find that I am more critical of lower (younger) grades, because I am unsure of appropriate knowledge of content at that age.
> Assessor 6, Site 4, Essay 3

While the NBPTS assessor training emphasizes that everyone has biases, and that there is no shame or fault in admitting this, many assessors still felt the need to reassure readers that their biases would not impede their work. There was a tendency among assessors to claim fairness

immediately after having admitted bias, which we interpret as a means to "save face" or reestablish professional status as a capable assessor. The following excerpts are taken from Essay 2 and Essay 3 responses in which an assessor, having made an explicit admission of bias, proceeds to provide the reader with a reassurance of fairness.

> I fully intend to *objectively gauge each entry* [emphasis added] here presented according to the set rubric and believe in the thought that teachers remain teachers because of the noble intent to produce thinking citizens for the future
> Assessor 1, Site 1, Essay 2

> The only thing I had to be especially careful about was letting the type of class (learning level or subject matter) affect my attitude … I can *assure the Board* and myself I *maintained a professional view* throughout [emphasis added]
> Assessor 3, Site 1, Essay 3

> A commentary that is well written and follows the National Guidelines *could affect my scoring*. I have tried not to be biased about this and instead look mainly for evidence given … I feel that I *can be a good assessor without any bias* [emphasis added].
> Assessor 1, Site 4, Essay 2

> *I might be hindered* because I came from a university that was very specific on their criteria for a competent teacher … *I am also very objective*, so having an open mind will help me to make a fair and accurate judgment [emphasis added].
> Assessor 6, Site 3, Essay 2

Such assertions of fairness immediately after an admission of bias occurred more often in Essays 2 and 3 than in Essay 1. As noted earlier, assessors were also more likely to make admissions of bias or give examples of bias in Essays 2 and 3 than in Essay 1. Together, these findings indicate that many assessors progressed from believing that fairness alone was sufficient, to recognizing the existence of biases while still assuring

the reader that fairness was the overriding factor in the scoring process. These results are reassuring and provide strong evidence of the success of the training program.

*Actions to be taken in the light of potential bias.* When responding to the question of what might help or hinder them in making fair and accurate judgments, assessors not only made admissions of bias and listed specific examples of biases; they also provided specific information about what actions might help them make fair and accurate judgments. Assessors made reference to the value of the bias training or supplied information about specific actions that they would or did take in order to minimize the effects of the biases on their scoring. It is important that assessors made this connection between identifying potential areas of bias and actions that they could take, given that the NBPTS assessor training relies on the active participation of assessors in the process of identifying and consciously screening out biases as they arise. Action is a key component of the NBPTS's strategy for ensuring fair and accurate scoring.

Table 5 illustrates the breakdown of two specific categories of actions that assessors supplied: actions specifically related to the bias training such as use of the trigger lists, and actions related to other aspects of training such as referring to rubrics or benchmark cases. The three columns in Table 1 do not sum to 50 assessors, since some assessors identified both types of actions, and not all assessors completed all three essays.

**Table 5**

***Actions Suggested by Assessors to Minimize the Impact of Bias***

| | Number of essays | | | |
|---|---|---|---|---|
| | Essay 1 | Essay 2 | Essay 3 | Total |
| No action suggested | 24 | 19 | 15 | 58 |
| Actions related to the bias-reduction training | 8 | 19 | 15 | 42 |
| Actions related to other aspects of training | 20 | 14 | 17 | 51 |

*Note.* The Total column sums to 151 rather than 139 since 12 essays contain references to both types of actions.

Five of the eight actions related to bias-reduction training suggested in Essay 1, and four of the 20 actions related to other aspects of training, came from assessors who indicated in their essays that they had scored before.

As the third and fourth rows of the table show, 42 essays (30%) and 51 essays (37%) cited actions related to assessors using trigger lists and other aspects of training, respectively, to minimize bias. While a total of 58 essays did not contain any actions that an assessor might take to minimize the impact of bias, an assessor may have noted an action in one essay but not the other two. Looking at the data by assessor, across the three essays, 21 assessors (42%) did not make any mention of the role that the trigger lists or bias training would have in helping them reduce the impact of bias. Similarly 19 assessors (48%) did not mention the other more general aspects of training in any of their three essays. However, collapsing across the two types of action that an assessor might suggest as an approach to "helping them make fair and accurate judgments" only 7 of the 50 assessors (14%) did not mention any type of action that they could take.

The following excerpts illustrate how assessors thought about and used their trigger lists. Assessors described how the process of writing down triggers, of rereading the trigger list, or of discussing scoring difficulties with the trainer helped to minimize bias.

> *Making written responses to my biases* [emphasis added] helped me make accurate judgments.
> Assessor 11, Site 2, Essay 3

> The trigger list was a huge help for me because it made me *look at my biases up front* [emphasis added].
> Assessor 5, Site 5, Essay 2

> The use of my trigger list keeps me on track.
> Assessor 3, Site 3, Essay 3

> The trigger list helped me make fair judgments. During the training we were to write down things that we felt would bias our judgment. *Referring back to this* [emphasis added] was very beneficial.
> Assessor 13, Site 2, Essay 3

> I had to *reread the trigger list* [emphasis added] to dismiss any + or – bias that could have interfered with accurate judging.
> Assessor 6, Site 1, Essay 3

> I *bounced them [biases] off the trainer* [emphasis added] when I felt they were significant.
> Assessor 5, Site 1, Essay 3

These quotations are in line with NBPTS's expectations concerning the role of the trigger lists (as part of the scoring materials), namely, to slow assessors down and give them time to consider their judgments in the light of potential bias areas, and to revise those judgments if necessary.

Similarly, as the following quotes illustrate, being mindful of an area of bias is the first step towards minimizing its impact.

> The whole process of going over bias was very helpful as it brought these to the forefront of my mind and awareness.
> Assessor 3, Site 1, Essay 3

> Bias training allowed me to incorporate newly found ideas about myself into more valid scoring opportunities.
> Assessor 4, Site 1, Essay 2

In addition to the benchmark cases, assessors made reference to other aspects of the training that they felt contributed to their ability to make fair and accurate judgments, citing aspects of the training such as standards, rubrics, and multiple practice examples. Forty percent of assessors indicated that the bias training was helpful, 18% made specific reference to their trigger lists, and 16% made reference to the benchmark cases.

The following examples illustrate how assessors perceived other aspects of the training to be useful.

> The many example cases we did has helped me make more accurate judgments.
> Assessor 2, Site 1, Essay 2

> The process of assessment becomes less bias [sic] as your perspective of what is accomplished and not accomplished becomes clearer.
> Assessor 1, Site 2, Essay 3

> I believe I can make a fair and accurate judgment of entries consistent with other assessors because of specific training that helped me to "internalize" a set of rubrics.
> Assessor 6, Site 1, Essay 2

> Rereading the rubric and any adjudication discussions have helped me further internalize the various ratings on the scale.
> Assessor 2, Site 1, Essay 3

> Being aware of triggers and following the rubric/note-taking guide as I followed the list of evaluation words guided me to accurately evaluate each entry.
> Assessor 3, Site 4, Essay 3

These previous comments tie in closely with training claims—that training consists of a two-strand approach, focusing on what is valued in the scoring process (as seen in the rubric and benchmarks) and making explicit what is not part of the scoring process (personal, societal, and writing biases).

One area of potential concern is that 34% of assessors identified some type of external influence that had the potential to hinder them from making fair and accurate judgments. These external issues included areas of physical discomfort such as being tired, cold, or hungry; or other distractions such as noise, interruptions, or being rushed.

In summary, a number of assessors supplied examples of specific actions that they could take to minimize the impact of bias, including such things as referring to their trigger list, consulting with the trainer, refreshing their memory of the benchmarks, or using other parts of the scoring apparatus (rubrics, benchmarks, etc.).

Initial analyses do not show any pattern of connections between assessors' supplying a specific example of bias and supplying a specific action that they could take to counteract it. Approximately one third of the assessors did not make any comments regarding using the trigger lists or the bias training to help minimize bias; one third of the assessors did not make any comments regarding the use of the more general scoring apparatus; and only seven assessors (14%) did not list any actions across all three essays.

### *Trigger Lists*

The second evidence source for this study was the assessors' trigger lists, copies of which were collected at the end of the4th day of training. Developing these personalized lists is an important aspect of the assessors' training. Assessors did not know that they would be asked to share them with the researchers at the end of the training, and they were given the option of not sharing their lists. The trigger lists provide a different type of insight into the assessors' thinking about bias during the training process. The lists are developed throughout the training, and additional entries may be added during scoring. The trainer never looks at or comments on the specifics of what assessors write—it is the assessors' own private view of their personal triggers.

The trigger lists vary in terms of the number of items that assessors identified and the degree of specificity of their entries. In the set of 31 trigger lists collected, assessors averaged 16 triggers, with a low of 7 and a high of 33.

The left column in Table 6 shows the 19 categories that were developed to classify the contents of the assessors' trigger lists. The underlined headings in the left column indicate the larger categories into which these 19 initial categories were collapsed. The right column provides a brief description of the type of triggers included in each category.

The five largest categories for the triggers were writing (23%), candidate character (12%), teaching (10%), candidate appearance (8%), and speech (7%). Given that these triggers were written in response to assessors' reactions based on video-based entries, it is not surprising that candidates' appearance and manner of speaking are two of the larger categories. Writing was the largest category, however, accounting for almost one quarter of all triggers listed.

As shown in Table 6, three of the categories of triggers are the same as those used in coding examples of bias in the essays: personal biases, writing biases, and societal biases. Writing triggers were grouped from the beginning as a distinct group and not further teased

apart. Although they initially formed the largest category, when examined at the more gross level, it becomes clear that personal biases heavily outweighed the writing biases.

**Table 6**

***Categories Used for Trigger Lists***

| Category title | References made to: |
|---|---|
| Personal bias | |
| Age | Age of candidate |
| Candidate appearance | Physical appearance of the candidate |
| Classroom appearance | Physical appearance of the classroom, e.g., neatness, disorganization |
| Behavior of students | Classroom management issues, behavior, noise level, students on/off task |
| Candidate character | Nonphysical characteristics, e.g. creative, confident, condescending |
| Emotional response | Feelings of sympathy/empathy, or interest generated by novel activities, etc. |
| Gender | Gender of candidate |
| Organization | Organization (or lack of) not specifically attributed to anything |
| Parent/community | Involvement of parents/community |
| School | Location of school—urban/suburban/rural |
| Speech | Candidate's accent, dialect, speaking skills |
| Students | Number of students in the classroom, large or small class size |
| Video | Quality of video |
| Writing bias | |
| Writing | Written commentary, e.g., presentation, neatness, grammar/spelling |
| Societal bias | |
| Race | Race, ethnicity of candidates or students |
| Socioeconomics | Socioeconomic differences/effects |
| Fatal flaws | |
| Teaching | Aspects of teaching that are connected to the rubric—goals, awareness (or lack) of what is going on in the classroom |
| Content | Content errors or knowledge demonstrated |
| External influences | |
| External | Effects that do not relate to any aspect of the candidate's submission such as assessor fatigue, time of day, being rushed |

A category not used in previous discussions emerged from the review of the trigger lists, a category referred to by the NBPTS training manual as "fatal flaws." After the trigger list is first introduced in the training, an overhead transparency is presented to assessors, listing potential triggers for their consideration. The first section of the overhead deals with fatal flaws, which are defined in the following way:

> This occurs when an assessor finds a critical "fault" in the response and allows this one error to discredit the other parts of the response. In fact, the scoring system is complex and holistic—no single error can "sink" a response. Types of fatal flaws may include:
> Content errors …Poor safety habits …Problematic Goals … Student Work … Student misbehavior … Lost opportunities …
> (NBPTS, 2002, page 32)

The final category, "external influences," includes items such as fatigue, temperature of the room, etc. These triggers were not counted in this analysis given that they were not part of a particular candidate's response.

Previous data did not lend itself to a breakdown by site, given the low numbers, but as can be seen in Figure 3, the percentage of triggers in each grouping by site (personal, societal, writing, and fatal flaws) was remarkably consistent across the sites. (There were no trigger list data from Site 1.) As can be seen from the figure, the personal bias category accounted for a small but significant set of triggers identified by assessors. Overall, personal bias triggers accounted for 63% of what assessors noted on the trigger lists, while writing bias triggers accounted for 23% and societal bias triggers for 4%. The category called "fatal flaws" accounted for 10%. In addition, at two sites, a small number of assessors listed potential triggers that were categorized as external influences.

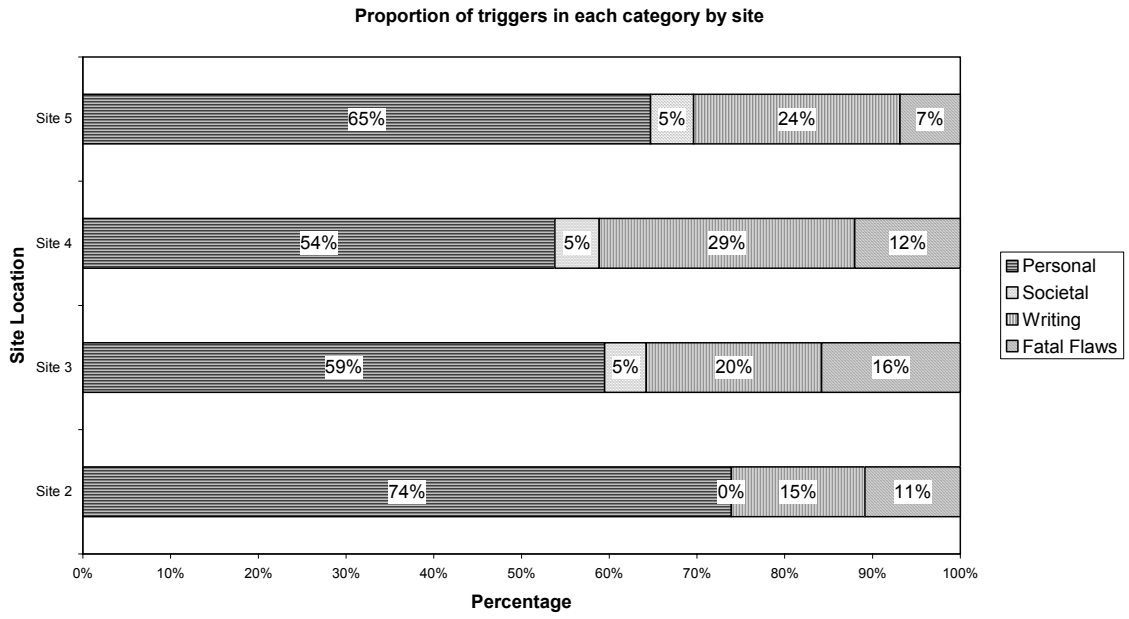


***Figure 3*. Bar chart illustrating the proportion of triggers in each category by site.**

### *Similarities in Findings Between Essays and Trigger Lists*

To examine changes in assessors' perceptions of themselves and their awareness of their biases, two major sources of data were used: assessors' essays and their "trigger lists." It was expected that assessors would incorporate vocabulary from the training sessions into both the trigger lists and the essays. The trigger lists were expected to contain specific words and phrases from the training, because the development of the trigger lists followed directly after particular bias-reduction exercises. Essays 2 and 3, written after training was completed, were also expected to contain words and phrases taken directly from the training.

After examining both sources of data in detail, the patterns uncovered were found to be strikingly similar. As shown in Table 7, in both the essay and the trigger list data, examples of personal bias were supplied most frequently, followed by writing biases. In both cases, very few instances of societal bias were provided. In the essays, societal biases were mentioned only six times (8%), compared to 49 times (64%) for personal biases and 22 times (29%) for writing biases. In the trigger lists, societal bias triggers accounted for 4% of triggers listed, personal bias triggers accounted for 63%, and writing bias triggers accounted for 23%. The fourth category, "fatal flaws," accounted for 10% of triggers listed.

**Table 7**

***A Comparison of Types of Bias Cited in Essays and Trigger Lists***

| Type of bias | Essays | Trigger lists |
|---|---|---|
| Personal | 64% | 63% |
| Writing | 29% | 23% |
| Societal | 8% | 4% |

Another area of comparison focused on the vagueness or lack of specificity in the assessors' entries. In the essays, assessors employed a variety of linguistic techniques to circumvent explicit admissions or discussions of bias. These included vocabulary choice ("feelings" versus "biases"), distancing through word choice or semantic arrangement, and assurances of fairness and objectivity.

The trigger lists differed from the assessor essays in that the majority of the comments were very specific. This was not surprising, because assessors were asked to identify specific issues that were likely to make them score a candidate higher or lower than the response deserved; thus, there was less opportunity to avoid explication of this potentially uncomfortable topic. Some assessors did manage, however, to provide some degree of circumvention by choosing vague words and phrases such as "dress," "appearance," or "tone," which did not indicate the type or direction of the bias (e.g., "professional, stylish dress," "neat and professional appearance," or "monotonous voice").

About 10% of all the triggers listed by assessors were classified as not being specific. The nonspecific triggers were related to appearance, speech and writing mostly. These nonspecific triggers were distributed fairly evenly, with most assessors having one or two of them. A notable exception was one assessor who had 8 nonspecific triggers out of a list of 25 (32%). Whether that assessor was reminded about a specific trigger on reading that list could not be ascertained from the current data.

Both of the data sources, the essays and the trigger lists, support the overall findings that the training appeared to assist assessors in identifying their biases, at least on a surface level. Assessors were either better able to or more comfortable in identifying personal and writing biases than societal biases. Because many assessors did mention the use of the trigger lists when

answering the essay prompts, it could be surmised that assessors actively employed the trigger lists as a tool to keep awareness of their biases within their conscious attention while scoring. This supports the Board's design for training and scoring and provides evidence that assessors were self-monitoring their biases.

## Conclusions

This study examined the NBPTS training program for assessors, focusing on the efficacy of the bias-awareness exercises. Data taken from assessors' trigger lists and from essays written at three points during the training and scoring process were used to evaluate assessors' perceptions of themselves and their awareness of their biases over the course of the training.

The study specifically examined the types of bias admissions the assessors made, the degree of explicitness of these admissions, and how these admissions changed over time. The trigger lists and essays were analyzed separately and then compared to determine if the two sets of data supported the same conclusions. If the assessors "applied" the training to themselves, as seen in their trigger lists and essays, then learning may have taken place. Learning in this case was defined as increasing awareness of biases in general, of the assessors' own biases in particular, and of the potential effects of bias on the work of the assessors.

Overall, the bias-awareness exercises appeared to produce the intended effect of increasing assessors' awareness of their own biases and the impact that those biases could have on the scoring process. This positive finding is tempered by evidence that this increased awareness was limited in scope and did not include a noticeable increase in discussions of societal bias.

Many assessors simply restated the words and phrases that were utilized in the training without individualizing the terms to their own personal biases, which might indicate a need for greater explicitness and more focus on having assessors apply the training to identify their individual biases. On the other hand, some assessors were able to build on and extend the training work on biases, providing evidence in support of the efficacy of the bias training. Because the evidence for these findings was limited to written texts (the trigger lists and essays) it is not known if more societal biases might have been revealed through other data collection methods, such as interviews with assessors.

Although the specific biases noted both in the assessors' essays and on trigger lists might vary somewhat across video entries (depending on the certificate area), the triggers noted appear to be consistent across the five observation sites. Student-work-based entries and documented accomplishments entries (the other entry types that make up the NBPTS portfolio) would certainly raise some different types of triggers since the assessors would not see the candidate, the students, or the classroom.

With regard to assessors' admissions of bias, assessors varied widely in the degree to which they took ownership of their biases, ranging from no admission of bias to explicit lists of biases. After training was completed, the majority of assessors made at least one explicit admission of bias (88%), a few only made implicit admissions (12%), and just one assessor out of 50 wrote all three essays without an explicit or implicit admission of bias.

Based on bias-awareness research and communications research, it was expected that the number and degree of explicit examples of bias would be greatest for personal bias, followed by writing bias, and lastly societal bias. The evidence found in both the trigger lists and the essays supported this expectation. Some assessors did not use the term "bias" at all in their essays, referring instead to "feelings," "reactions," or "personal likes and dislikes." Some assessors used the term "personal bias," one assessor used the phrase "writing bias," and no assessors used the phrase "societal bias."

Across the 139 essays written by assessors, 78 (56%) essays did not contain any specific examples of bias. In the remaining 61 essays, personal bias examples were supplied 49 times (64%), writing biases 22 times (29%), and societal biases six times (8%). Overall, assessors did not supply specific examples of bias as frequently as they made admissions of bias.

Given that there were fewer specific examples of bias than admissions of bias, no strong patterns regarding changes in assessors' examples of bias over time emerged from the data. For a limited number of assessors, there was some degree of progression in explicitness and/or number of biases mentioned across the three essays.

On the other hand, one third of the assessors made assertions of fairness and objectivity based on their past teaching experiences, and another third asserted that they could be fair because of their open-mindedness. The majority of these comments occurred in the first of the three essays, with more explicit admissions of bias appearing in the second and third essays. The finding that most assertions of fairness were replaced with admissions of bias and/or examples of

bias in the second and third essays provides evidence of the success of the training program in increasing assessors' awareness of their personal biases.

In addition to assertions of fairness, assessors employed a range of linguistic techniques to avoid direct ownership of their biases. This aligns with research on racial awareness (Derman-Sparks & Phillips, 1997; Sleeter 1995) and findings by discourse analysis researchers on avoidance of socially taboo topics in public forums (Rosenberg, 1997; van Dijk, 1987). The linguistic techniques used to provide distance between the assessor and any bias included the use of second or third person pronouns, the absence of any person reference in the sentence, the use of adverbs for emphasis on surprise in finding bias, and the use of indirect sentence structure to create distance between the actor and the bias.

Given that the National Board relies on the active participation of assessors in the process of identifying and consciously screening out biases as they arise, it was important that assessors both admit their biases and take action on them. Fifty-eight percent of the assessors mentioned one or more specific actions to take while scoring that related to the use of the trigger lists or discussing biases with the trainer. As expected, three fourths of those suggestions were made in Essay 2 and Essay 3, after training had been completed.

A broader category of "actions" was created to include not only referring to the trigger list and consulting with the trainer, but also reviewing benchmarks or using other parts of the scoring apparatus (rubrics, note-taking guide, etc.). Using this composite category of actions taken, 86% of assessors indicated an action that they could or would take to ensure fair and accurate scoring. Only 14% of assessors did not mention specific actions they could take to minimize bias while scoring. However, the essay prompts did not specifically ask for actions, only "what would help or hinder" them in making fair and accurate judgments.

The validity of the NBPTS assessment process rests largely on the structured professional judgments of trained assessors, applying scoring rubrics in a fair and consistent manner. The bias-awareness training that assessors receive is designed to bring any previously unarticulated influences to the surface, and by doing so, to minimize their effects on scoring. This study provides valuable confirmation of the efficacy of the bias-awareness exercises. The study also suggests that further work could be done to guide assessors in specifically examining societal biases and to emphasize the importance of taking conscious action to minimize the impact of their biases during scoring.

## References


Acosta-Deprez, V. M. (1995). Approaches to making the comprehensive school health education curriculum multicultural. In J. M. Larkin & C. E. Sleeter (Eds.), *Developing multicultural teacher education curricula* (pp. 229-46). Albany, NY: State University of New York Press.

Banks, J. A. (1995). Multicultural education: Its effects on students' ethnic and gender role attitudes. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 617-627). New York: Macmillan.

Banks, J. A. (2001). *Cultural diversity and education: Foundations, curriculum, and teaching*. Boston: Allyn and Bacon.

Baron, R., Tom, D., & Cooper, H. (1985). Social class, race, and teacher expectations. In J. B. Dusek, V. C. Hall, & W. J. Meyer (Eds.), *Teacher expectancies* (pp. 251-270). Hillsdale, NJ: Erlbaum.

Bernadin, J. H., & Beatty, R. W. (1984). *The process of performance appraisal: Assessing human behavior at work.* Boston, MA: Kent Publishing Company.

Bogdan, R.C., & Biklen, S. K. (1992). *Qualitative research methods for education* (2nd ed.) Boston: Allyn and Bacon.

Brief of Amicus Curiae American Psychological Association, Gratz v. Bollinger & Grutter v. Bollinger, (6th Cir. 2003) (No. 02-241 & No. 02-516, available at http://www.psyclaw.org/grutter-v-bollinger.pdf).

Brophy, J., & Everston, C. (1981). *Student characteristics and teaching*. New York: Longman.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Delpit, L. (1995). *Other people's children: Cultural conflict in the classroom*. New York: New Press.

Derman-Sparks, L., & Phillips, C. B. (1997). *Teaching/learning anti-racism: A developmental approach.* New York: Teachers College Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5-18.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York: Macmillan.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*(2), 127-148.

Freedman, P., Gotti, M., & Holtz, G. (1983). A comparison of the effectiveness of two approaches to the reduction of stereotypical thinking. *Contemporary Education*, *54*(2), 134-138.

Giroux, H. (1998). Youth, memory work, and the racial politics of Whiteness. In J. L. Kincheloe, S. Steinberg, N. Rodriguez, & R. Chennault (Eds.), *White reign: Deploying whiteness in America* (pp. 123-136). New York: St. Martin's Press.

Grant, C., & Secada, W. (1990). Preparing teachers for diversity. In W. Houston (Ed.), *Handbook of research on teacher education* (pp. 403-422). New York: Macmillan.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology 74*(6), 1464-1480.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*(6), 1022-1038.Gumperz, J. (1999). On interactional sociolinguistic method. In C. Roberts & S. Sarangi (Eds.), *Talk, work and institutional order: Discourse in medical, mediation and management settings* (pp. 453-471). Berlin: Mouton de Gruyter.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge, England: Cambridge University Press.

Guttmann, J., & Bar-Tal, D. (1982). Stereotypic perceptions of teachers. *American Educational Research Journal*, *19*(4), 519-528.

Hale-Benson, J. (1982). *Black children: Their roots, culture, and learning styles.* Baltimore: Johns Hopkins University Press.

Helms, J. E. (Ed.). (1990). *Black and white racial identity: Theory, research, and practice*. New York: Greenwood Press.

Hymes, D. (1962). The ethnography of speaking. In T. Gladwin & W. C. Sturtevant (Eds.), *Anthropology and human behavior* (pp. 13-53). Washington, DC: Anthropological Society of Washington.

Ilgen, D. R., & Feldman, J. M. (1983). Poor performers; Supervisors' and subordinates' responses. *Organizational Behavior and Human Performance*, *27*, 386-410.

Julian, M. A. (1996, May 18). *Influences, issues, and trends in educating the educators on the other side of the Atlantic*. Paper presented at the Penn-TESOL East Spring Conference, University of Delaware, Newark, DE.

Khera, N. (1995). *P.A.C.E. (Programs for Awareness in Cultural Education): The study of a peer education program in cross-cultural awareness at the University of Pennsylvania*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Lewis, A. E. (2001). There is no "race" in the schoolyard: Color-blind ideology in an (almost) all-white school. *American Educational Research Journal*, *38*(4), 781-811.

Martin, R. J. (Ed.). (1995). *Practicing what we teach: Confronting diversity in teacher education.* Albany, NY: State University of New York Press.

Michalski, R. S. (1993). Beyond prototypes and frames: The two-tiered concept representation. In I. Van Mechelen, J. Hampton, R.S. Michalski, & P. Theuns, (Eds.), *Categories and concepts: Theoretical views and inductive data analysis*. London: Academic Press.

National Board for Professional Teaching Standards. (2002). NBPTS portfolio assessor training manual certificates 18-64. Arlington, VA: Author.

Pate, G. S. (1995). *Prejudice reduction and the findings of research*. University of Arizona School of Education: Information Analyses.

Pohan, C. A., & Aguilar, T. E. (2001). Measuring educators' beliefs about diversity in personal and professional contexts. *American Educational Research Journal*, *38(*1), 159-182.

Rosenberg, P. M. (1997). Underground discourses: Exploring Whiteness in teacher education. In M. Fine, L. Weis, L. C. Powell, & L. M. Wong (Eds.), *Off White: Readings on race, power, and society* (pp. 79-89). New York: Routledge.

Sadker, M., Sadker, D., & Long, L. (1993). Gender and educational equity. In J. A. Banks & C. A. Banks (Eds.), *Multicultural education: Issues and perspectives* (2nd ed., pp. 111-128). Boston: Allyn & Bacon.

Schiffrin, D. (1994). *Approaches to discourse*. Cambridge: Blackwell.

Schön, D. (1987). *Educating the reflective practitioner.* San Francisco, CA: Jossey-Bass.

Scollon, R., & Scollon, S. W. (2001). Intercultural communication, 2nd ed. Oxford, UK: Blackwell.

Shultz, K., Buck, P., & Niesz, T. (2000). Democratizing conversations: Racialized talk in a post-desegregated middle school. *American Educational Research Journal*, *37*(1), 33-65.

Slavin, R. E. (1995). Cooperative learning and intergroup relations. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 628-634). New York: Macmillan.

Sleeter, C. (1995). Teaching Whites about racism. In R. J. Martin (Ed.), *Practicing what we teach: Confronting diversity in teacher education* (pp. 117-130). Albany: State University of New York Press.

Szpara, M. (1999). *Talk among student teachers in an urban high school: Questioning dimensions of difference.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Thomas, D. R. (2003, August). *A general inductive approach for qualitative data analysis.*Retrieved March 18, 2004, from The School of Population Health, University of Auckland Web site: http://www.health.auckland.ac.nz/courses/ComHlth710/Inductive2003.pdf

van Dijk, T. (1984). *Prejudice in discourse: An analysis of ethnic prejudice in cognition and conversation.* Philadelphia: John Benjamins.

van Dijk, T. A. (1987). *Communicating racism: Ethnic prejudice in thought and talk.* Newbury Park, CA: Sage.

## Notes

[1] NBPTS Standards have been written for each certificate area and define what accomplished teachers should know and be able to do. The assessment is then based solely on the content of these standards.

[2] The Note-Taking Guide and the Scoring Path are documents that guide assessors in the note-taking process and the approach to scoring to be used on every case scored to ensure that all responses are treated in an identical manner. Along with the scoring rubrics, these three documents form the scoring apparatus used by NBPTS assessors.